Out[1]:

Out[2]: Click here to toggle on/off the Python code.

PHYS 10792: Introduction to Data Science

2019-2020 Academic Year

Course instructors: Rene Breton (http://www.renebreton.org) - Twitter @BretonRene (https://twitter.com/BretonRene)
Marco Gersabeck (http://www.hep.manchester.ac.uk/u/gersabec) - Twitter @MarcoGersabeck (https://twitter.com/MarcoGersabeck)

Chapter 10

Syllabus

1. Probabilities and interpretations
2. Probability distributions
3. Parameter estimation
4. Maximum likelihood + extended maximum likelihood
5. Least square, chi2, correlations
6. Monte Carlo basics
7. Probability
8. Hypothesis testing
9. Confidence level
10. **Goodness of fit tests**
11. Limit setting
12. Introduction to multivariate analysis techniques

Topics

10.1 Introduction

Many experiments extract their results by fitting a parametrisation of the observables to the data distribution. While fits can generally be convinced to converge and thereby give an answer the question arises of how trustworthy such results are.

This section will cover different approaches for assessing the goodness of fits. The fits themselves were discussed in week 5. Essentially, goodness of fit tests can be interpreted as hypothesis tests as well. The null hypothesis is that the fit model describes the data well and the test evaluates the significance of that agreement. A small significance would then lead to the hypothesis being rejected, i.e. the statement that the fit is bad.

10.2 Chi-squared test

10.2.1 General formulae

The χ^2 test is the most widely used goodness of fit test. In general, for sets of measurements x_i and y_i, where x_i is known precisely and y_i is measured with uncertainty σ_i, the χ^2 value is defined for the fit function $f(x_i)$ as

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_i^2}.$$

This function shows that each measurement will contribute roughly 1 to the total χ^2 assuming that the measured values fluctuate around the function values with a spread corresponding to their σ. Hence, χ^2 is expected to take a value of roughly N; however, we will discuss this more accurately below.

More generally, one can write the χ^2 formula as a matrix equation, which includes correlations between measurements, as

$$\chi^2 = (\mathbf{y} - \mathbf{f})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{f}),$$

where $\mathbf{V}$ is the covariance matrix of the measurements.

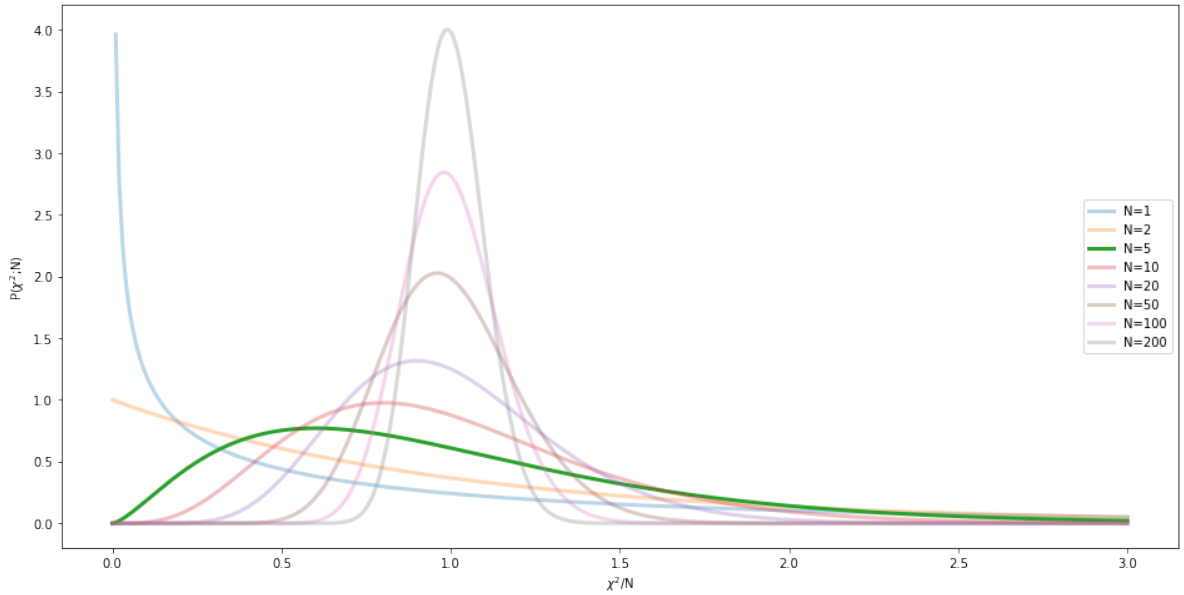
The probability distribution for χ^2 is given by

$$P(\chi^2; N) = \frac{2^{-N/2}}{\Gamma(N/2)} \chi^{N-2} e^{-\chi^2/2},$$

which can be used to calculate the significance of the test as

$$\mathrm{Prob}(\chi^2; N) = \int_{\chi^2}^{\infty} P(\chi'^2; N) d\chi'^2.$$

This is called the χ^2 probability. Let's have a look at it:



This plot shows χ^2 probability density functions for various degrees of freedom (N). For ease of comparison the x axis has been scaled by N.

We can see that the mean indeed is approximately the number of degrees of freedom. However, the variance of the distribution differs strongly with N.

Over to you: menti.com (https://www.menti.com) with key 88 09 86.

One often comes across statements like "$\chi^2/N < 1.5$ indicates a good fit". This is rather pointless as the significance of this statement varies strongly with N. The following shows the probabilities of a fluctuation of the null hypothesis exceeding this threshold for different degrees of freedom.

```
  N Prob chi2>1.5N
  1 0.220671
  2 0.223130
  5 0.186030
 10 0.132062
 20 0.069854
 50 0.012597
100 0.000904
200 0.000006
```

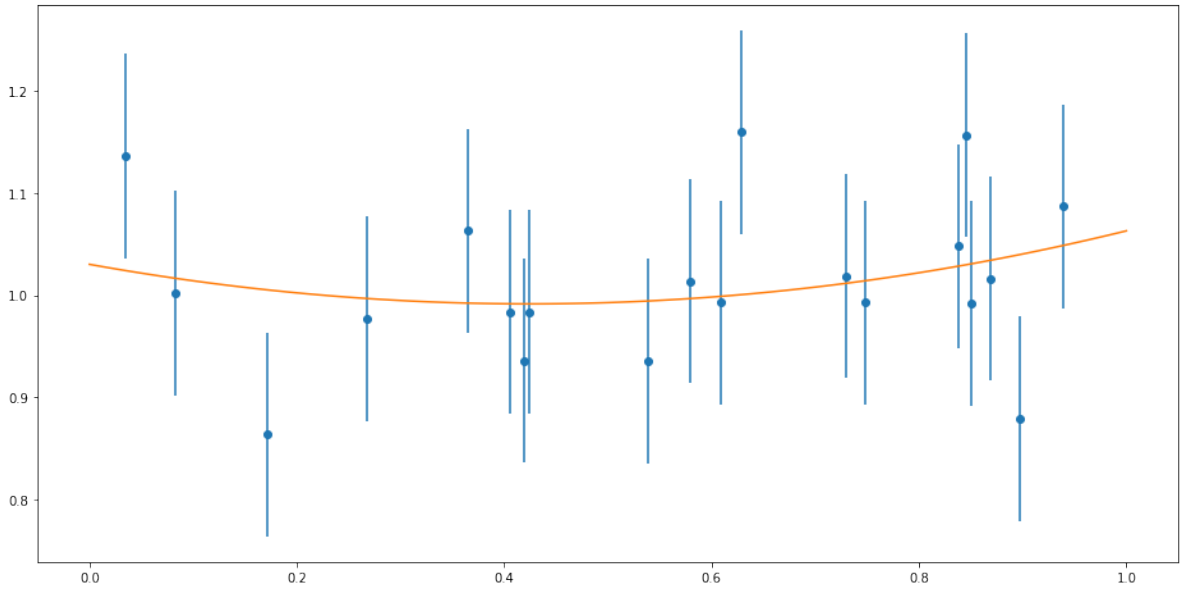
So rather than fixing a useless limit in χ^2/N, we should fix the desired significance and deduce the corresponding acceptable limit in χ^2 (or χ^2/N). So let's have a look how this pans out for a limit of 1%.

```
  N chi^2/N
  1 6.634897
  2 4.605170
  5 3.017254
 10 2.320925
 20 1.878312
 50 1.523078
100 1.358067
200 1.247226
```

Example: fit to un-binned dataset

Having worked out how to set an acceptance threshold we can now have a look at how this applies to a concrete dataset. In this case we need to account for the correct number of degrees of freedom. These are in principle the number of measurements, but they are reduced by each of the free parameters of the fit function. This is because each of the fit parameters is optimised to minimise the χ^2 during the fitting process. If any of the function parameters are not fitted but rather fixed to a specific value, the number of degrees of freedom should *not* be reduced.

In the example below we have data points ranomly distributed between 0 and 1 with y values that are randomly distributed around 1. As a fit function we use a parabola. As this function has 3 parameters, we use as the number of degrees of freedom the number of measurement points minus 3.



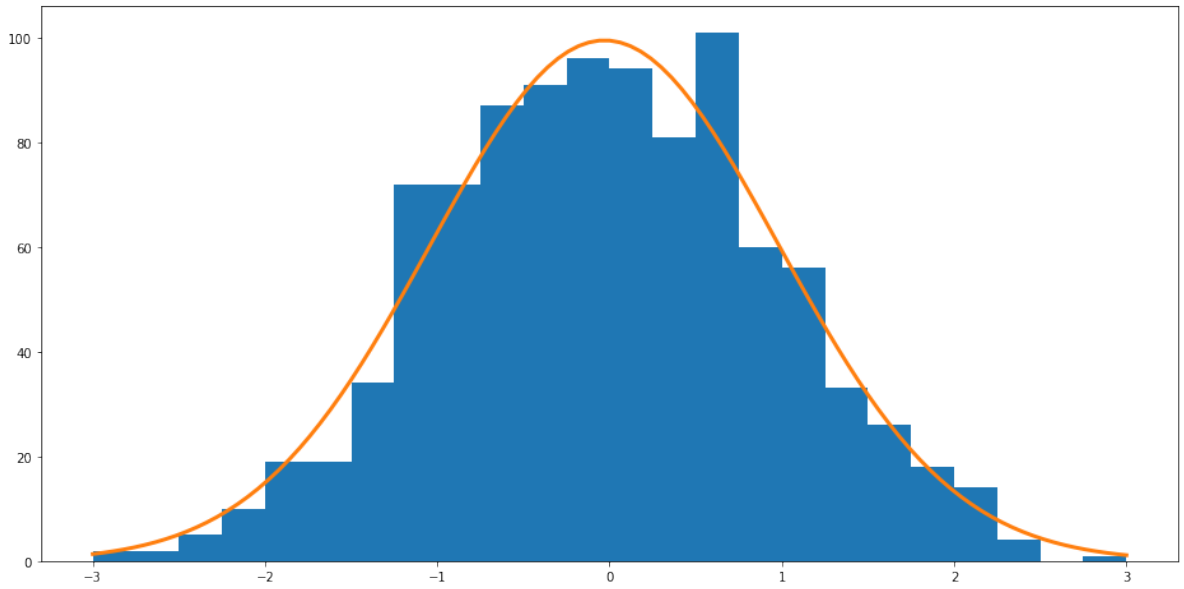

```
Fit results:
Par. 0: 0.99151 +/- 0.02811
Par. 1: -0.21537 +/- 0.26754
Par. 2: 0.42392 +/- 0.18909
The fit chi2 is 11.7 for 20 measurements and 3 fit parameters, i.e. 17 degrees
of freedom.
The corresponding probability is Prob(11.7,17)=82.01663%.
```

Example: fit to binned dataset

In principle things are very similar for a binned dataset, i.e. when the fitted data are represented in a histogram. The main difference is that the number of degrees of freedom is no longer given by the number of elements in the dataset but rather by the number of bins of the histogram. The reduction by the number of fit parameters is unchanged.

A second difference is the uncertainty used in the χ^2 calculation. This is no longer the uncertainty of the individual entries contributing to the histogram, but the uncertainty on the cumulative content of each bin. For a counting experiment with sufficiently many counts the uncertainty would be the square root of the number of counts in each bin.

So let's have a look how this pans out in an example. We generate a dataset that is drawn from a normal distribution and fit this with a Gaussian function.




```
Fit results:
Par. 0: 252.29999 +/- 8.50126
Par. 1: -0.02987 +/- 0.03936
Par. 2: 1.01226 +/- 0.03944
The fit chi2 is 22.9 for 24 bins and 3 fit parameters,i.e. 21 degrees of freedo
m.
The corresponding probability is Prob(22.9,21)=34.79233%.
```

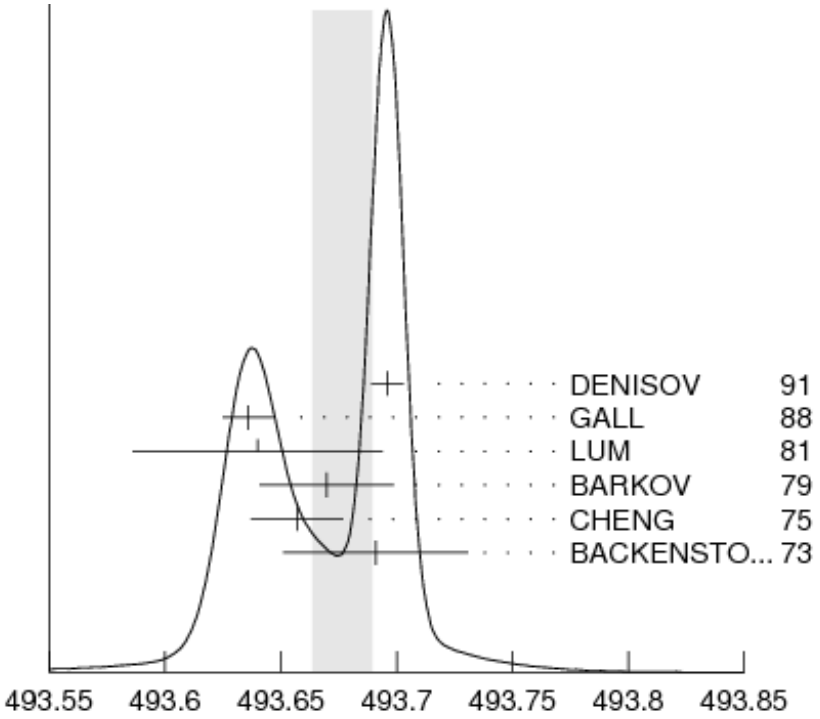
As we know the expected shape we can also fix the parameters of the function and test the agreement of this with the data. In this case the number of degrees of freedom is not reduced by the number of parameters.

```
The fit chi2 is 20.6 for 24 bins,i.e. 24 degrees of freedom.
The corresponding probability is Prob(20.6,24)=66.01792%.
```

10.2.2 Rescaling χ^2 distributions

As we have seen, the distribution of χ^2/N has a mean of roughly 1, independent of N. In some cases when several measurements are being combined to extract their average or a more complicated derived result one can be confronted with the situation that the measurements do not agree particularly well according to a χ^2 test.

The plot below shows several measurements of the $K^{\pm}$ meson mass in units of MeV/c^2. It is apparent that they do not agree perfectly. The grey band indicates the world average extracted from these measurements.



Source: C. Patrigiani et al. (Particle Data Group), 2017 Review of Particle Physics, Chin. Phys. C, 40 (2016) 100001 and 2017 update (http://pdglive.lbl.gov/DataBlock.action?node=S010M&init=0)

Your call now what to do with this:

menti.com (https://menti.com) with key 88 09 86.

If none of the input measurements can be dismissed and all uncertainties appear to have been assessed correctly, one is faced with two choices:

- Either one performs the combination as planned and just notes the tension between the measurements without doing anything about it,
- Or one inflates the uncertainty of the combination to account for the poor level of agreement of the inputs.

In the latter case the question arises by how much to inflate the uncertainty. A commonly used approach in this case is to scale the input uncertainties by exactly the amount that is required to bring the reduced χ^2 to $\chi^2/N = 1$. This is achieved by the following scaling

$$\sigma^{\text{scaled}} = \sqrt{\chi^2/N}\sigma^{\text{measured}}.$$

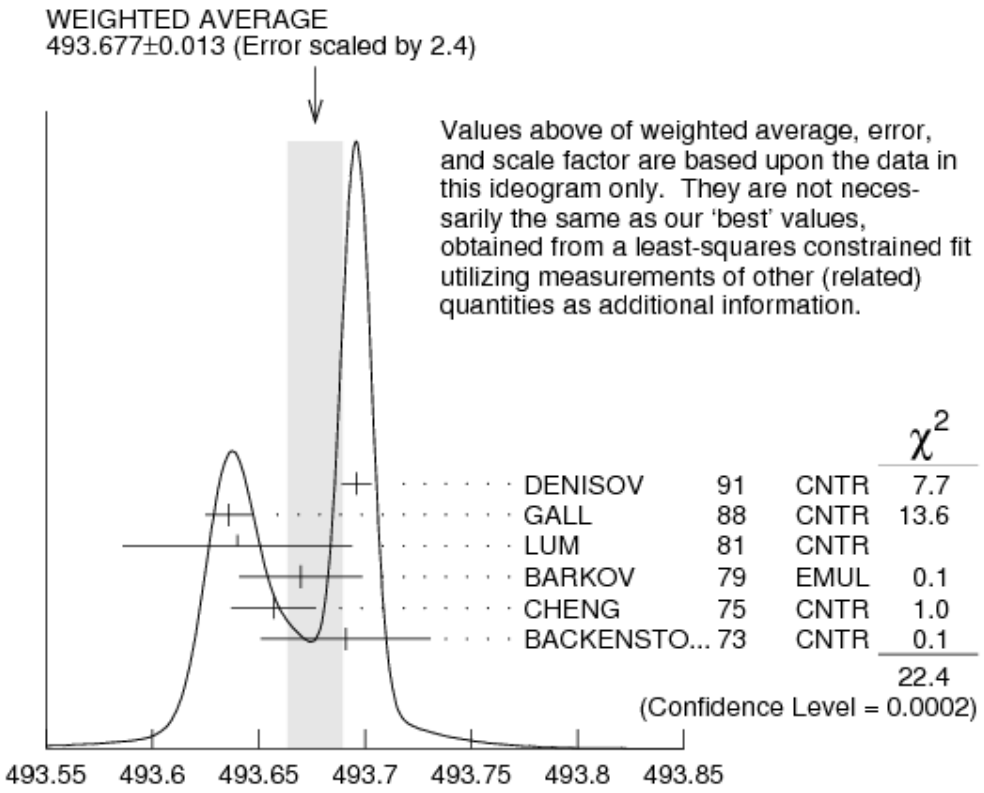
This procedure relies on the two assumptions that:

- the model is correct (this is particularly relevant when fitting a more complex model to the data to extract an underlying parameter rather than simply calculating a mean), and that
- only the scaling of the data uncertainties is unknown (i.e. there is no other conceivable way of reconciling the measurements).

It is worth noting that this scaling does not change the central value of the average.

The Particle Data Group (PDG) applies this scaling to their averages. They scrutinise all candidate input measurements and discard any that are deemed to be inappropriate for inclusion in the average. This can be simply because the measurement has been superseded by another one that is included or because it is considered to be of inferior quality for some reason.

The application of this procedure to the $K^{\pm}$ mass looks as follows.



You may note that the scale factor of 2.4 corresponds to $\sqrt{22.4/4}$. This is because the PDG exclude measurements with too large uncertainties from the calculation of their scaling factors as they might affect the χ^2 significantly without actually having much impact on the average. In this case, this is the case for *Lum 81* and so the number of degrees of freedom is $5-1=4$, where the reduction of 1 is to account for the average that is a fit parameter.

10.3 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test is used for comparing a distribution of data events to a given distribution. This distribution needs to be fully defined rather than fitted to the data as the KS test has no means of taking into account the reduction in the number of degrees of freedom.

The KS test is based on constructing cumulative distributions both of your data, $\mathrm{cum}(x)$, and of the function to be compared to the data, $\mathrm{cum}(P)$,. Both should be normalised such that the cumulative distributions go from 0 to 1. You then need to find the largest difference between the two cumulative distributions, i.e.

$$D = max|\mathrm{cum}(x) - \mathrm{cum}(P)|.$$

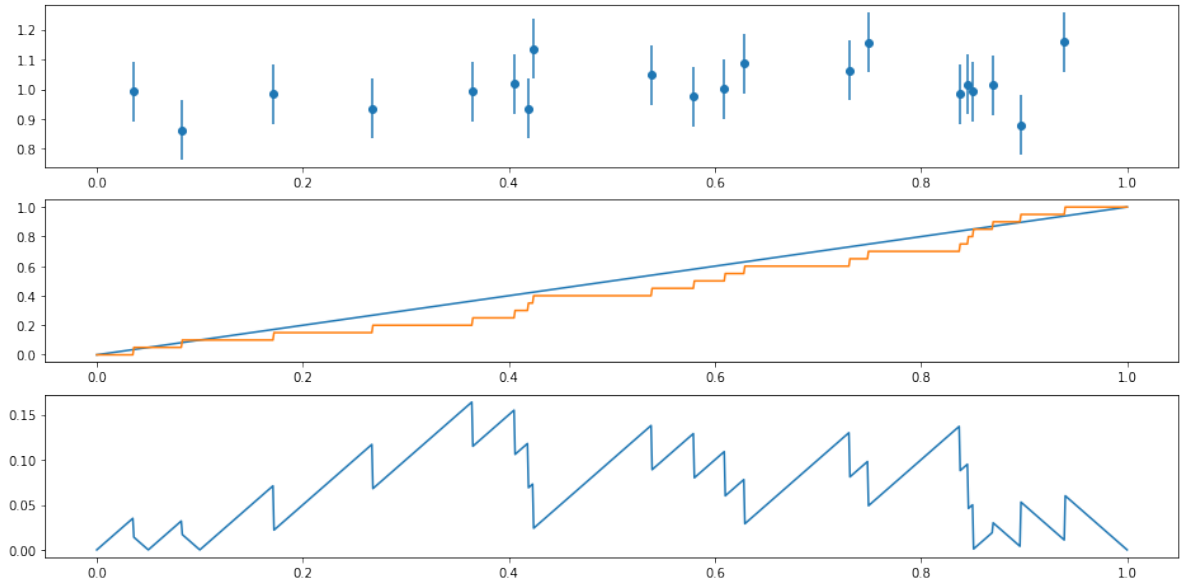
Finally, we need to account for the sample size as this drives the agreement of the two distributions. Hence, we scale D and introduce

$$d = D\sqrt{N}.$$

The value of d then needs to be compared to a table of critical values, c, to determine the level, α, beyond which the hypothesis that both distributions are compatible is rejected, i.e. you require $d < c(\alpha)$.

c (α)	α
1.63	0.01
1.36	0.05
1.22	0.10
1.07	0.20

We can now apply this to a dataset. Let's choose the x values of the unbinned dataset above. They should be drawn from a uniform distribution between 0 and 1, so the cumulative distribution, $\mathrm{cum}(P)$, is a straight line. The cumulative distribution of the data, $\mathrm{cum}(x)$, is a step function that increases by $1/N$ at the position of each data point, where N is the total number of data points in the set.



```
The KS test output for 20 entries is D=0.164 and d=0.733.
```

10.4 Two-sample problem

We often encounter situations where, rather than comparing a dataset to a fit curve, we need to compare two datasets. The questions asked can vary, from the comparison of a single parameter describing an aspect of their shape, e.g. mean or width, to a general comparison of whether two distributions agree with the hypothesis of being drawn from a single parent distribution.

Whatever the exact question that is being looked at, what we are dealing with is a hypothesis test with the null hypothesis being that the two shapes or particular aspects thereof are in agreement with being drawn from a common parent distribution.

Comparing samples with known σ

One encounters quite frequently the scenario that two measurement outcomes have to be compared and checked whether they are compatible with each other. Let us assume here that the uncertainty of the measurements is known.

You measure results x_1 and x_2 and want to compare whether they are compatible. Their resolutions are σ_1 and σ_2.

This is equivalent to testing the hypothesis that their difference,

$$x_1 - x_2,$$

is compatible with zero.

The variance of the difference is

$$V_{12} = \sigma_1^2 + \sigma_2^2,$$

so essetially we want to compare the difference, $x_1 - x_2$, to the combined uncertainty $\sigma_{12} = \sqrt{V_{12}}$.

There are two examples where this applies.

Example 1: Two measurements with different known resolution

Suppose you conduct a lab experiment in pairs and you and your partner apply different methods to measure the same quantity.

As an example, let's assume you measure the height of a building. One method is a free-fall experiment where you drop a stone down the side of the building (with all necessary safety precautions) and measure the time it takes to reach the bottom. The second method is triangulation from a point nearby.

The first method yields a height of (25.4 ± 0.4) m, and the second (24.0 ± 0.3) m.

The difference of the measurements is $1.4m$ and the combined uncertainty is $0.5m$, so the discrepancy is at the level of 2.8 standard deviations. This corresponds to a confidence level of 99.49% (see e.g. Barlow Table 3.2 or the calculation below).

```
99.49%
```

If this happened to you in first-year lab, would you reject right away the hypothesis that the two measurements are compatible?

menti.com (https://menti.com) with key 74 58 08.

With about 300 students there are 150 pairs who all conduct several experiments throughout the year. The discrepancy of 2.8σ has a chance of occurring of about 0.5% or 1 in 200. Therefore this *should* happen every now and then.

In *any* case, and not just when stumbling over a potential discrepancy, you should always question your results and check whether you have accounted for all possible **systematic uncertainties**. In this particular example this could be for example the effect of friction or the reaction time for the free-fall experiment, or the accuracy of the height that you are extrapolating from in the triangulation (the list goes on).

Example 2: Two ensembles of measurements

Another case is where we have two ensembles with a known spread and we want to test the compatibility of their means. An example of this scenario is your exams, which should neither be too easy nor too hard.

Let's assume one course has 310 students taking the exam and an average grade of 65.8%. Another course has 55 students who achieve on average 72.3%. Both distributions have a spread of 8%.

The difference to assess is 6.5%.

The uncertainty on the mean of the first course is $8\%/\sqrt{310} = 0.45\%$ and for the second one it is $8\%/\sqrt{55} = 1.08\%$. Hence, the uncertainty on the difference is 1.17%.

The discrepancy is over 5 standard deviations and is therefore statistically significant.

However, also here **systematic uncertainties** apply. For example, the smaller course might have attracted students who, on average, tend to do better in exams. It is also not possible to set exams with the level of predictability suggested by the $1/\sqrt{N}$ scaling of the large courses.

Nevertheless, while the numbers here are made up, the rest of the discussion is definitely part of reality and exam outcomes are analysed very carefully and appropriate actions are taken where required.

When σ is unknown

In the previous example we assumed a spread of 8% for the marks. Unless this was doctored in some way, the only way to obtain this number would be by measuring the spread from the distribution itself. This means that we are not using an unbiased estimator of $\boldsymbol{\sigma}$.

For the numbers of students discussed in the example we get away with that. For smaller sample sizes we would have to refer to the Student's t distribution to take this fact into account properly.

The calculation gets somewhat more involved albeit not prohibitively so, but is beyond the scope what we can cover here. If you encounter a situation like that where numbers are small, make sure to read up on the proper handling of this. The Student's t distribution was mentioned in week 2 and is discussed e.g. in chapters 7.3 and 8.4.2 of Barlow.

10.5 Kolmogorov-Smirnov test with two samples

The KS test can also be applied to the two-sample problem. Instead of comparing a distribution of discrete events with a continuous distribution one can also compare two discrete distributions.

In this case one creates two staircase-like cumulative distributions, $\mathrm{cum}(x)$ and $\mathrm{cum}(y)$, that cover the two samples of N_x and N_y events. Once more, we need to find the largest difference between these distributions,

$$D = \max|\mathrm{cum}(x) - \mathrm{cum}(y)|.$$

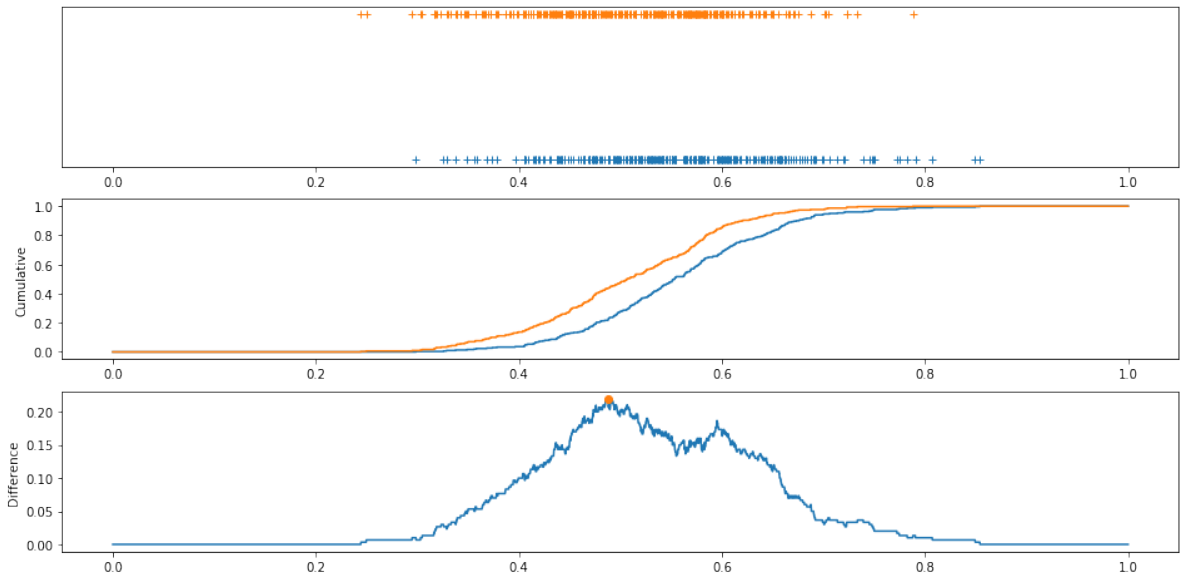
In this case, the appropriate scaling is achieved by

$$d = \sqrt{\frac{N_x N_y}{N_x + N_y}} D.$$

The critical values are the same as those given before and are repeated here for convenience.

c (α)	α
1.63	0.01
1.36	0.05
1.22	0.10
1.07	0.20

The following example compares two distributions that are drawn from a Gaussian distribution with $\mu = 0.5, \sigma = 0.1$, and $N_x = 20, N_y = 300$.




```
The KS test output for 300 and 300 entries is D=0.220 and d=2.694.
```